D622–D627 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj083

dbPTM: an information repository of protein post-translational modification

Tzong-Yi Lee[1], Hsien-Da Huang[1,2,*], Jui-Hung Hung[1], Hsi-Yuan Huang[1], Yuh-Shyong Yang[2,3] and Tzu-Hao Wang[4]

[1]Institute of Bioinformatics, National Chiao Tung University, Hsin-Chu 300, Taiwan, [2]Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan, [3]Institute of Biochemical Engineering, National Chiao Tung University, Hsin-Chu 300, Taiwan and [4]Department of Obstetrics and Gynecology, Chang Gung Memorial Hospital, Lin-Kou Medical Center, Tao-Yuan 333, Taiwan

Received August 15, 2005; Revised and Accepted October 11, 2005

ABSTRACT

dbPTM is a database that compiles information on protein post-translational modifications (PTMs), such as the catalytic sites, solvent accessibility of amino acid residues, protein secondary and tertiary structures, protein domains and protein variations. The database includes all of the experimentally validated PTM sites from Swiss-Prot, PhosphoELM and O-GLYCBASE. Only a small fraction of Swiss-Prot proteins are annotated with experimentally verified PTM. Although the Swiss-Prot provides rich information about the PTM, other structural properties and functional information of proteins are also essential for elucidating protein mechanisms. The dbPTM systematically identifies three major types of protein PTM (phosphorylation, glycosylation and sulfation) sites against Swiss-Prot proteins by refining our previously developed prediction tool, KinasePhos (http://kinasephos.mbc.nctu.edu.tw/). Solvent accessibility and secondary structure of residues are also computationally predicted and are mapped to the PTM sites. The resource is now freely available at http://dbPTM.mbc.nctu.edu.tw/.

INTRODUCTION

Protein post-translational modification (PTM) is an extremely important cellular control mechanism because it may alter proteins' physical and chemical properties, folding, conformation, distribution, stability, activity and consequently, their functions (1). Examples of the biological effects of protein modifications include phosphorylation for signal transduction, attachment of fatty acids for membrane anchoring and association, and glycosylation for changing protein half-life, targeting substrates, and promoting cell-cell and cell-matrix interactions. With the accelerating progress in proteomics, biological knowledgebases containing a wealth of information, in particular protein modifications, are playing crucial roles in cell regulation research (2).

The Swiss-Prot knowledge base (3) includes as much modification information as is available with consistency and structure, allowing easy retrieval by biologists. Phospho.ELM (1), which was developed as part of the ELM (Eukaryotic Linear Motif) resource, is a new resource containing experimentally verified phosphorylation sites that were manually curated from the literature. O-GLYCBASE (4) is a database of glycoproteins, most of which include experimentally verified O-linked glycosylation sites. The RESID protein modification database is a comprehensive collection of annotations and structures for protein modifications and cross-links including pre-, co- and post-translational modifications (5). The RESID database provides modification information, literature citations, Gene Ontology (GO) cross-references, protein sequence database feature table annotations, structure diagrams and molecular models. Each RESID entry presents a protein with a chemically unique modification and indicates how the modification is currently annotated in the Swiss-Prot (6).

In this study, we collect the known PTM information from external biological data sources. Since only a small fraction of Swiss-Prot proteins are annotated with experimentally verified PTMs, we also developed computational tools to comprehensively identify phosphorylation sites, glycosylation sites and sulfation sites against the Swiss-Prot proteins. Protein structural properties and functional information, such as the solvent accessibility of residues, protein variations, non-synonymous

*To whom correspondence should be addressed. Tel: +886 3 5712121, ext. 56952; Email: bryan@mail.nctu.edu.tw
Correspondence may also be addressed to Tzu-Hao Wang. Tel: +886 3 3281200, ext. 8984; Email: knoxtn@cgmh.org.tw

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

single nucleotide polymorphism (SNP), protein tertiary structures and protein functional domains, are provided for researchers who are investigating the protein PTM mechanisms. Web query interface and graphical visualization were designed and implemented to facilitate access to the database content.

DATA GENERATION

The data generation flow of the dbPTM is briefly depicted in Figure 1. The data generation flow comprises the three major components: integration of external known PTM databases, learning and prediction of PTM sites, and structural or functional annotations. The experimentally validated PTM data sources were extracted from Swiss-Prot (3), Phospho.ELM (1) and O-GLYCBASE (4). The experimentally verified PTM sites were used to generate computer models to further identify putative PTM sites against the Swiss-Prot proteins. Additional structural properties and functional information, such as protein tertiary structures, protein secondary structures, solvent accessibility of residues, protein functional domains, protein variations and non-synonymous SNP, are also annotated to the Swiss-Prot proteins. The detailed data generation flow is described below.

Integration of external known PTM databases

Three external biological databases related to protein PTM information, Swiss-Prot (3), Phospho.ELM (1) and O-GLYCBASE (4), are integrated into the proposed resource. Both the experimentally validated PTM sites and the putative PTM sites, which are annotated as 'by similarity', 'potential' or 'probable' in the 'MOD_RES' fields, have been extracted from the Swiss-Prot database (3). As summarized in Table 1, release 46.0 of Swiss-Prot contributes 11 025 experimental validated PTM sites within 4921 proteins, and 72 308 putative PTM sites within 31 026 proteins. The Phospho.ELM entries store information about substrate proteins with the exact positions of residues are known to be phosphorylated by cellular kinases. A total of 1703 experimentally verified phosphorylation sites within 556 proteins were obtained from Phospho.ELM version 2 (1). O-GLYCBASE (4) Version 6.00 provides 242 glycoproteins containing 2765 experimentally

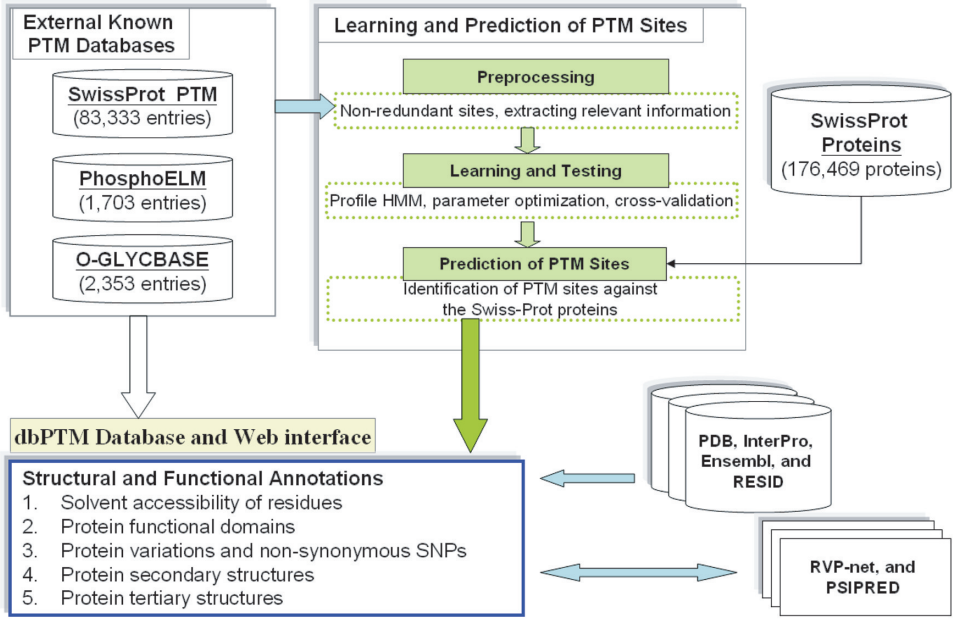

Figure 1. The data generation flow of the dbPTM database.

Table 1. The list of the integrated external data sources

Databases	Description	Statistics
Swiss-Prot (3,12)	Experimental PTMs	11 025 PTM sites within 4921 proteins
	Putative PTMs	72 308 PTM sites within 31 026 proteins
	Proteins	176 469 Proteins
	Protein variants	32 101 Variants corresponding to 6115 proteins
PhosphoELM (1)	Experimental phosphorylation sites	1703 PTM sites within 556 proteins
O-GLYCBASE (4)	Experimental glycosylation sites	2353 PTM sites within 185 glycoproteins
RESID (5)	PTMs	373 PTM types
InterPro (14)	Protein domain	1 113 928 Entries corresponding to 161 988 Swiss-Prot entries
Ensembl (13)	Human variations	23 378 Non-synonymous SNPs within 7230 Swiss-Prot human proteins
PDB (15)	Protein structures	31 721 Entries corresponding to 6806 Swiss-Prot proteins

verified O-linked, N-linked and C-linked glycosylation sites. Moreover, 185 glycoproteins in O-GLYCBASE are corresponded to Swiss-Prot proteins, which have 2353 experimentally verified glycosylation sites.

Learning and prediction of PTM sites

To provide the PTM information of the PTM un-annotated proteins available from Swiss-Prot, we integrated several computational tools for identifying the PTMs of the Swiss-Prot proteins. Our previous work, namely KinasePhos (7), incorporated the profile hidden Markov model (HMM) to identify kinase-specific phosphorylation sites with ~87% prediction accuracy (8), which was compared with several phosphorylation prediction tools, such as NetPhos (9), DISPHOS (10) and rBPNN (11) (see Supplementary Table S1). KinasePhos is integrated and used to fully detect the kinase-specific phosphorylation sites against the Swiss-Prot proteins. To reduce the number of false positive predictions by KinasePhos, we set the predictive parameters as the values when the prediction specificity is 100% (8).

As depicted in Supplementary Figure S1, the KinasePhos-like method, which is similar to KinasePhos for phosphorylation sites, was designed and implemented to learn models for the prediction of the sulfation sites, N-linked glycosylation sites and C-linked glycosylation sites (Table 2). We used the 144 known sulfation sites of tyrosine, 1790 N-linked glycosylation sites of asparagine and 49 C-linked glycosylation sites of tryptophan to evaluate the performance of the KinasePhos-like prediction tools. The result suggests that all three KinasePhos-like tools exhibited high prediction precision, sensitivity and specificity (Supplementary Table S2).

Structural and functional annotations

In order to provide more effective information about protein structural and functional annotations relevant to protein PTM, a variety of biological databases, such as Swiss-Prot (12), Ensembl (13), InterPro (14), PDB (15) and RESID (5), are integrated.

Protein variation is the change of amino acids in polypeptides. As summarized in Table 1, Swiss-Prot contributes 32 101 protein variants corresponding to 6115 proteins, where 47 variant residues are located at the PTM sites and 267 variant residues are located surrounding 236 PTM sites ($-4 \sim +4$ AA). Furthermore, single amino acid polymorphism (SAP) is the amino acid variation corresponding to the genetic variation as the definition of non-synonymous SNP in genomic sequence. The amino acid variants may have an impact on protein folding, active sites, or the overall solubility and stability of a protein. SAP is the type of variation most frequently related to human diseases (12). Therefore, when the amino acid variations occur in the PTM sites or the surrounding residues, they may affect the recognition of PTM sites by catalytic kinases. A total of 23 378 human non-synonymous SNPs located at 7230 Swiss-Prot human proteins were obtained from the variation part of Ensembl database (13).

InterPro provides 1 113 928 entries corresponding to 161 988 Swiss-Prot proteins. We found that about 65% of Swiss-Prot annotated PTM sites are located at InterPro annotated protein domains. The RESID (5) protein modifications database is integrated into dbPTM to provide PTM related information, such as mass difference, chemical formula, enzymatic activities, literature citations, GO cross-references, structure diagrams and molecular models.

The latest version of PDB contains 31 721 tertiary structures corresponding to 6806 Swiss-Prot protein entries (Table 1). For the proteins with known tertiary structures, the DSSP (16) program was used to extract the true secondary structure and solvent accessibility for those 6808 Swiss-Prot proteins. Solvent accessibility of amino acids residues is important for both the structure and function of proteins, especially the PTMs studied in this investigation. Protein secondary structure is the regular arrangement of amino acid residues in a segment of a polypeptide chain, where each amino acid is assigned a structure state, helix (H), strand (E) or coil (C). There are 1124 experimentally verified PTMs have the true secondary structure and solvent accessibility.

However, only ~4% of Swiss-Prot proteins have the known tertiary structures. For proteins without known tertiary structures, two previously published tools, RVP-net (17) and PSIPRED (18), were applied to predict the solvent accessibility and the secondary structure, respectively (see Table 2). RVP-net (17) presents a feed-forward type neural network which can predict a real value ranging from 0 to 100% of accessible surface areas (ASAs) for amino acid residues, based on their neighborhood information. We applied the RVP-net program (17) to fully predict the real-valued ASA for the amino acid residues of all Swiss-Prot proteins. By selecting a suggested threshold (17) (i.e. 25%), the residues with larger ASA values are viewed as surface residues.

Table 2. The list of the integrated annotated tools

Tools	Description
KinasePhos (7)	Identifying kinase-specific phosphorylation sites
KinasePhos-like sulfation	Identifying sulfation sites
KinasePhos-like N-linked glycosylation	Identifying N-linked glycosylation sites
KinasePhos-like C-linked glycosylation	Identifying C-linked glycosylation sites
DSSP (16)	Calculating the secondary structure and solvent accessibility of residues
RVP-net (17)	Predicting the solvent accessibility of residues
PSIPRED (18)	Predicting the protein secondary structures
Weblogo (15)	Generating sequence logo for PTM substrates

DATA STATISTICS

The statistics of the experimentally verified PTMs and the putative PTMs compiled in the dbPTM resource are shown in the Table 3. For instance, dbPTM contains 14 057 known PTM sites and 772 154 putative PTM sites. The parameters of the predictive tools, KinasePhos, KinasePhos-like Sulfation and KinasePhos-like Glycosylation—for the prediction of phosphorylation sites, sulfation sites and glycosylation sites, respectively—are set as the values when the predictive specificity is set to 100% during the parameter optimization of the trained models (8). The numbers of putative phosphorylation and sulfation sites, where the ASA of the substrates are >25% (defined as the residue locating at the protein surface), are 652 756 and 13 315, respectively. There are a total of 33 887 predicted N-linked glycosylations of asparagine and C-linked glycosylations of tryptophan.

Table 3. The data statistics of the dbPTM database

PTM types	Substrates	No. of known PTMs	No. of putative PTMs	Total
Phosphorylation	Serine, threonine, tyrosine, aspartate, histidine or cysteine	3367	5852	
	Serine, threonine and tyrosine (predicted in this resource, ASA > 0%)		1 346 067	661 975 (ASA > 25%)
	Serine, threonine and tyrosine (predicted in this resource, ASA > 25%)		652 756	
Glycosylation	N-linked, O-linked and C-linked glycosylation	4586	55 059	
	N-linked asparagines and C-linked tryptophane (predicted in this resource, ASA > 0%)		43 894	94 132 (ASA > 25%)
	N-linked asparagines and C-linked tryptophane (predicted in this resource, ASA > 25%)		33 887	
Sulfation	Serine, threonine and tyrosine	144	413	
	Tyrosine (predicted in this resource, ASA > 0%)		189 457	13 872 (ASA > 25%)
	Tyrosine (predicted in this resource, ASA > 25%)		13 315	
Lipidation	GPI-anchor, N-terminal myristoylation and palmitoylation	520	4688	5208
Acetylation	N-terminal of some residues and side chain of lysine or cysteine	1019	1580	2599
Amidation	Generally at the C-terminal of a mature active peptide after oxidative cleavage of last glycine	1554	523	2077
Methylation	Generally of N-terminal phenylalanine, side chain of lysine, arginine, histidine, ralinenes or glutamate and C-terminal cysteine	455	1105	1560
Hydroxylation	Generally of ralinenes, aspartate, raline or lysine	816	515	1331
Pyrrolidone carboxylic acid	N-terminal glutamine which has formed an internal cyclic lactam	567	408	975
Gamma-carboxyglutamic acid	4-Carboxyglutamate	343	263	606
Trimethylation	N6-methylated lysine, N6,N6,N6-trimethyllysine, N,N,N-trimethylalanine	158	294	452
Blocked	Unidentified N- or C-terminal blocking group	108	10	118
FAD	O-8alpha-FAD tyrosine, Pros-8alpha-FAD histidine, S-8alpha-FAD cysteine and Tele-8alpha-FAD histidine	12	77	89
S-nitrosylation	S-nitrosocysteine	5	59	64
Formylation	Of the N-terminal methionine	35	27	62
Deamidation	Deamidated asparagin and deamidated glutamine (needs to be followed by a G)	33	18	51
Citrullination	Citrulline	7	41	48
Others		328	1274	1134
Total		14 057	772 154	786 211

ASA, accessible surface area.

INTERFACE

To facilitate the use of the dbPTM resource, we developed a website for users to browse and search for content. As depicted in Supplementary Figure S2, the user can select a particular type of PTM for browsing the information. When clicking on a PTM entry, it pops up a window showing the solvent accessibility of the residues, the secondary structures and the flanking sequence of the PTM site.

The search pages allow users to query the database using the Swiss-Prot ID and protein name. The interface also presents structural properties and functional information corresponding to the resulting proteins, such as the solvent accessibility of residues, non-synonymous variations, protein domains and protein secondary structures. Furthermore, the positional relationships among the PTMs, protein structural properties and protein functional information are graphically displayed (Figure 2).

Generally, a 3D presentation is an effective manner for revealing the PTM information corresponding to the protein tertiary structures. For these purposes, we developed a protein structure viewer for the visualization of protein tertiary structures and especially of the port-translational modification residues. As shown in Supplementary Figure S3, the visualization tool provides a comprehensive view of the whole protein structure and marks residues that are annotated as the PTM sites. This visualization tool is implemented as a client-side tool based on OpenGL's pipeline.

The visualization of the protein structures and the annotated residues are provided by two different ways according to different users' platforms. For users in MS Windows, the users can download the installable package of the Silver. After the Silver is installed, the protein tertiary structures and the PTM sites can be graphically and directly provided, as shown in Supplementary Figure S3. Alternatively, for users in other platforms such as Mac OS X, Linux and Solaris, the user can download the PDB structure and the Rasmol (http://www.umass.edu/microbio/rasmol/) scripts for the labeling of the PTM sites.

CONCLUSIONS

The proposed resource not only integrates the experimentally validated PTM information, but it also computationally annotates the Swiss-Prot proteins for putative phosphorylation, glycosylation and sulfation sites. Furthermore, the PTM related protein structural properties and functional information, such as solvent accessibility of amino acid residues, protein variations, protein secondary structures, protein tertiary structures and protein domains, are provided to facilitate the research of protein PTMs.

Figure 2. The graphical interface reveals the PTMs, the solvent accessibility of the residues, protein variations, protein secondary structures and protein functional domains.

One of the prospective goals for dbPTM is to integrate more efficient prediction tools for other types of PTM in addition to phosphorylation, sulfation and N- and C-linked glycosylation. Other protein sequence databases besides the Swiss-Prot protein database can also be considered and annotated for post-translation modifications by the proposed resource.

AVAILABILITY

The dbPTM resource will be regularly maintained and updated. The resource is now freely available at http://dbPTM.mbc.nctu.edu.tw/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Shih-Yee M. Wang (University of Illinois at Chicago) for English editing, the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 94-2213-E-009-025 (to H.-D.H.), and Chang Gung Memorial Hospital for the Research Grant CTRP1006 (to T.-H.W.). Funding to pay the Open Access publication charges for this article was provided by Chang Gung Memorial Hospital.

Conflict of interest statement. None declared.

REFERENCES

1. Diella,F., Cameron,S., Gemund,C., Linding,R., Via,A., Kuster,B., Sicheritz-Ponten,T., Blom,N. and Gibson,T.J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*,, **5**, 79.
2. Farriol-Mathis,N., Garavelli,J.S., Boeckmann,B., Duvaud,S., Gasteiger,E., Gateau,A., Veuthey,A.L. and Bairoch,A. (2004) Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics*, **4**, 1537–1550.
3. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
4. Gupta,R., Birch,H., Rapacki,K., Brunak,S. and Hansen,J.E. (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res*, **27**, 370–372.
5. Garavelli,J.S. (2004) The RESID database of protein modifications as a resource and annotation tool. *Proteomics*, **4**, 1527–1533.
6. Wu,C.H., Yeh,L.S., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al*. (2003) The Protein Information Resource. *Nucleic Acids Res*., **31**, 345–347.
7. Huang,H.D., Lee,T.Y., Tzeng,S.W. and Horng,J.T. (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res*., **33**, W226–W229.
8. Huang,H.D., Lee,T.Y., Tzeng,S.W., Wu,L.C., Horng,J.T., Tsou,A.P. and Huang,K.T. (2005) Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. *J. Comput. Chem*., **26**, 1032–1041.
9. Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol*., **294**, 1351–1362.
10. Iakoucheva,L.M., Radivojac,P., Brown,C.J., O'Connor,T.R., Sikes,J.G., Obradovic,Z. and Dunker,A.K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*., **32**, 1037–1049.

11. Berry,E.A., Dalby,A.R. and Yang,Z.R. (2004) Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms. *Comput. Biol. Chem.*, **28**, 75–85.
12. Yip,Y.L., Scheib,H., Diemand,A.V., Gattiker,A., Famiglietti,L.M., Gasteiger,E. and Bairoch,A. (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat.*, **23**, 464–470.
13. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
14. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P., Bucher,P. *et al.* (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform.*, **3**, 225–235.
15. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
16. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
17. Ahmad,S., Gromiha,M.M. and Sarai,A. (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics*, **19**, 1849–1851.
18. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.